DOCUMENT RESUME

ED 317 571 TM 014 612

AUTHOR Chan, Jason C.
TITLE Response-Order Effect in Likert-Type Scales.
PUB DATE 18 Feb 90
NOTE 31p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Testing; Correlation; Estimation (Mathematics); Factor Analysis; *Factor Structure; Foreign Countries; High Schools; *High School Students; Item Analysis; Item Response Theory; *Likert Scales; *Primacy Effect; Student Reaction; *Test Format; Test Items
IDENTIFIERS *Item Position (Tests); *Response Patterns; Taiwan (Taipei)

ABSTRACT

The importance of the presentation order of items on Likert-type scales was studied. It was proposed that subjects tend to choose the first alternative acceptable to them from among the response categories, so that a primacy effect can be predicted. The effects of reversing the order of the response scale on the latent factor structure underlying the items and the item and person parameters estimated by item response theory were also studied. Five Likert-type items, theoretically constituting a Personal Distress Scale, were answered by 102 senior high school students (49 males and 53 females) in Taipei City (Taiwan). Five weeks later, the same students responded to the scale with the ordering of the scale labels reversed. Primacy effects were supported, and the items' underlying factor structures were changed regardless of whether the factor analysis was based on Pearson or polychoric correlations. Estimated item parameters indicated that items with a traditional response scale order (from positive to negative) were "easier" than were those with reversed order. Persons whose parameters were estimated from items in traditional response scale order had higher trait levels than those estimated from items with the reversed order. Thus, primacy effects interfered with the item response theory theoretical property of parameter invariance. Practical implications of the findings are discussed. Five tables present study data, and a 44-item list of references is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED317571

RESPONSE-ORDER EFFECT IN LIKERT-TYPE SCALES

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JASON CHIHYU CHAN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Jason C. Chan

Department of Educational Psychology

University of Texas at Austin

Austin, TX 78712

2/18/1990

Running Head: Response-Order Effects

Postal Address:

Educational Building 504

Department of Educational Psychology

The University of Texas at Austin

Austin, TX 78712

* The author wants to thank Dr. William R. Koch for his helpful critiques and rhetoric suggestions.

TM 014012

BEST COPY AVAILABLE

Abstract

The present study argues that the meaning of verbal labels of a Likert-type response scale is affected by the presentation order of the scale labels, namely the response order. It is proposed that subjects tend to choose the first alternative acceptable to them from among the ordered response categories so that a primacy effect is predicted. In addition, the effects of reversing the ordering of the response scale on (a) the latent factor structure underlying the items and (b) the item and person parameters estimated by Item Response Theory (IRT) are assessed. A total of 102 high school students responded to 5 Likert-type items which theoretically constituted a Personal Distress Scale. Five weeks later, they responded to the same items with the ordering of the scale labels reversed. Results showed that primacy effects were supported and that the items' underlying factor structures were changed regardless of whether the factor analysis was based on Pearson or polychoric correlations. Item parameters estimated from the Graded Response Model of IRT indicated that items with a traditional response scale order (i. e., from positive to negative) were "easier" than those with reversed order. Consequently, person whose parameters were estimated from the items with the traditional response scale order had "higher" trait levels than those estimated from the items with the reversed order. Thus, primacy effects interfered with IRT's theoretical property of parameter invariance. Practical implications of the response-order effects were also discussed.

Response-Order Effects in Likert-Type Scales

Likert-type scales are widely used in market research, social survey, and personality assessment. However, the total variance in responses to an attitude scale is determined by the whole measurement environment; such basic factors as differences in stimulus objects, subjects' characteristics, and measurement instrument characteristics may have a systematic impact on subjects' responses (Sauber, 1983). Based on a meta-analysis, Churchill and Peter (1984) concluded that measurement instrument characteristics have a much greater influence on reliability estimates than do sampling characteristics and measurement instrument development processes. The present study focuses on one of the measurement characteristics: the presentation order of Likert-type scales.

Context Effects

Ordinarily, the response scale for Likert-items is presented with the most positive descriptor first (e.g.: strongly agree, agree, neutral, disagree, strongly disagree). If the presentation order is reversed (e.g.: strongly disagree, disagree, neutral, agree, strongly agree), will the subjects' responses be affected? From the viewpoint that those scale labels should have standard, objective, and context-free meaning, the answer is "No!" This view is implicitly or explicitly assumed by many researchers as in the case where negatively worded items have the response scale flipped prior to data analysis. Some researchers have even tried

to assign fixed ranking or numerical values to those frequently used scale quantifiers. For example, Hakel (1968) presented a ranking of adjectives reflecting an increasing degree of frequency of occurrence of an unspecified event without controlling any context. Bass, Cascio, and O'connor (1974) also tried to associate "objective" quantities of frequency and amount with subjective modifiers. They established geometric means and standard deviations for 39 expressions of frequency and 44 expressions of amount. This assumption that scale modifiers have context-free meanings, however, has been seriously challenged (e.g., Bradburn and Miles, 1978).

The meaning of the verbal label of a scale may depend upon the context of the label. The first context to be considered is the position of the label. Chase (1969) suggested that the meaning of the scale adjectives is determined by the adjective's relative position in a group of response categories rather than the "standard" definition of the scale labels. This suggestion is consistent with Wildt and Mazis' (1978) findings that both label and location have an impact on subject response. Klockars and Yamagishi (1988) also found that the meaning of the labeled position is defined as a compromise between the label itself and the relative position. The second context to be considered is the nature of the event in the item stem (Parducci, 1968). For example, Pepper and Prytulak (1974) found that the mean and variance of an expression's numerical definitions depend upon both the expression and its context in which events' frequencies

varied. The final context to be considered is the set within which a descriptor is nested (Schertzer and Kerman, 1985).

Researchers from the area of cognitive psychology also suggest that the meanings of words are affected by contexts (Barsalou, 1982; Hintzman, 1986; Ortony, Schallert, Reynolds, & Anton, 1978; Roth & Shoben, 1983). Consequently, as stated by Worcester and Burns (1975), "the problem is not just that different words mean different things but that the same word can be made to mean different things as the context changes." The present study, therefore, argues against the traditional view that quantifiers/descriptors have context-free meanings; and it predicts that reversing the presentation order of verbal scales will change the meaning of the label and the labeled position which, in turn, will affect the subject's response. However, the direction and magnitude of the change of the subject's response needs to be clarified further.

Serial Position Effects

Reversing the order of response labels implies not only changing the position of the scale label but also reversing the sequence of subjects' information-processing. In this situation, a primacy or a recency effect is frequently considered. The current study focuses on the serial position effect of response options, although the effect also occurs for the order/position of items (e.g. Bishop, Oldendick, & Tuchfarber, 1985; Knowles, 1988).

As early as 1929, Mathews found that subjects were more likely to endorse response options printed on the left side of the page than on the right side. This primacy effect is also found when respondents are asked to make preference choices from a long list of political candidates (Brook & Upton, 1974; Mueller, 1970), from a list of 16 radio programs (Becker, 1954), or from a set of children's qualities (Krognick & Alwin, 1987). Belson (1966) asked his subjects to rate 36 questions about television on five kinds of intensity scales: satisfaction, agreement, approval, interest, and liking. He found a clear primacy effect: The categories at the positive end of scales in the traditional order (favorable to unfavorable) or at the negative end of scales in reversed order received greater endorsement when they were presented first than when presented last. This effect was replicated by Payne (1972) in a postal survey, although Payne claimed that the effect was smaller. In regard to evaluative items, Carp (1974) found that either the positive or negative end of the scale drew more responses when it was presented first. In contrast, some researchers have also reported recency effects, i.e., a response is more likely to be selected when it is given last or near the bottom of a list. Those recency effects tend to occur for dichotomous items (McClendon, 1986; Payne, 1951; Schuman & Presser, 1981) or for response options presented in oral form (Rugg & Cantril, 1944). Finally, a few researchers have failed to obtain clear evidence for serial position effects of response options (Campbell & Mohr, 1950; Johnson, 1981; Powers, Morrow, Goudy, & Keith, 1977).

The contradictory results discussed may be due to researchers' focusing on the empirical phenomenon without considering a theoretical framework to explain the psychological processes that do or do not produce the serial position effect. Recently, Tourangeau and Rasinski (1988) tried to develop a four-stage response process in attitude measurement. Specifically focusing on the response-order effect, Krosnick and Alwin (1987) also developed a theoretical model to explain the underlying processes. According to the latter model, Simon's (1957) satisfying principle is relevant for visually presented options. People usually respond with the first satisfactory or acceptable option in order to minimize cognitive costs rather than perform an exhaustive search for optimal solutions. Consequently, when the list of options is long and when a number of options are similar or equally appealing, a primacy effect is predicted. Orally presented options tend to produce recency effects because the capacity of short-term auditory memory is limited.

The present study employs 5-point intensity scales presented in written form. Since it is usually difficult for subjects to differentiate adjacent categories (e.g., slightly strong agreement versus moderately strong agreement) in finely anchored scales, a primacy effect is predicted.

Form-Resistant Psychometric Parameters

Even though it is known that the form of response scales has some effects on responses, a "form-resistant correlation" hypothesis (Stouffer & DeVinney, 1949) is usually assumed. Such a

hypothesis proposes that the form of response scales may have a constant effect on every subject who responds to the same form so that the inter-item correlations should not change due to changes in the form. This situation is similar to the homogeneity of variance assumption in the t-test or ANOVA that separate variance estimates provided by treatment group(s) and the control group are estimates of the same population variance. In other words, treatments are assumed to have a constant effect on every subject within the same treatment group so that the group's variance is not increased or decreased, although its mean may be changed. Kalton and Schuman (1982) have recognized that it seems dangerous to place uncritical reliance on the "form-resistant correlation" assumption. Krosnick and Alwin (1987; 1988) have tried to test this hypothesis empirically for different form effects. They found that changes in response order altered both the variances and the covariances of the items. However, they also found that the loadings of the items on the latent factor and the variance of the factor were invariant with regard to item/response order.

Factor loadings are estimates of the relationships between the latent trait and the manifest responses to the items. If one wishes to test the degree to which person and item parameters are invariant, Item Response Theory (IRT) is very appropriate for scales measuring a single dimension. When an IRT model fits item response data, the person parameters, i.e., the estimated levels of the latent trait, should be invariant regardless of the items to which responses are made. In addition, the item parameters, such as difficulty/threshold and discrimination parameters,

should also be invariant regardless of what samples of persons are used for calibration. These properties are known as "item-free person measurement" and "sample-free test calibration" (Wright, 1967). Since most researchers using factor analysis are interested in the underlying factor structure of an instrument, while most psychometricians using IRT are interested in person and item parameter estimates, the present study will examine the impacts of reversing the response scale order on both factor structures estimated from factor analysis and person and item parameters estimated from IRT.

Previous studies have usually compared two independent samples who responded to different orders of scales. Since those samples were not randomly selected, results from those studies are less valid than results from repeated samples. In the current study, repeated samples will be compared. Mean differences will be examined first. A primacy effect for response order is predicted. If this prediction is supported, the assumption that scale quantifiers have a context-free meaning will, in turn, be rejected. In addition, the factor structures underlying the two correlation matrices for two forms of response scales will be compared. The hypothesis of form-resistant factor structures implies that the two correlation matrices will be fit equally well with the same unidimensional model given that the quantities of the corresponding path coefficients are constrained to be equal. Finally, the effects of response order on person and item parameters estimated from an IRT model--Samejima's (1969) graded response model--will be investigated. The graded response model

is an extension of the two-parameter logistic model for dichotomously-scored items to the testing situation where item responses are scored into two or more ordered categories. The model's applicability to Likert-type scales has been demonstrated by several studies (e. g. Dodd, 1984; Koch, 1983; Thissen & Steinberg, 1988). The present study predicts that response order effects will produce item bias. More specifically, it is predicted that the difficulty parameters for each response category of each item will not be invariant because individuals with the same latent trait level do not have equal probability of choosing the same category of the same item due to response order effects.

Method

Subjects

A total of 102 students (49 males, 53 females) from two senior high schools in Taipei city participated in this study as part of the activities for counseling and guidance classes. Their ages ranged from 15 to 17.

Instruments

Five items from the Personal Distress (PD) Scale, a subscale of the Interpersonal Reactivity Index (Davis, 1980), were employed. They were: 1) "When I see someone who badly needs help in an emergency, I go to pieces;" 2) "I sometimes feel helpless when I am in the middle of a very emotional situation;" 3) "In emergency situations, I feel apprehensive and ill-at-ease;" 4) "Being in a tense emotional situation scares me;" and

5) "I tend to lose control during emergencies." Those five items were translated into Mandarin by Chan (1986). Each item was accompanied by a 5-point Likert-type response scale.

Procedures

In the first administration of the five items, the response scale labels were from the positive end to the negative end (from left to right) as follows: "describes me very well (4)," "describes me quite well (3)," "describes me well (2)," "describes me slightly well (1)," and "does not describe me well (0)." The scale form with this response order will be called the "positive form." Five weeks later, in the repeated administration of the same test, the above response order was reversed. The latter form will be called the "reversed form."

Results

Mean Differences

The common scoring procedure for Likert-type scales is simply to assign successive integers to response categories and sum over the items to obtain the total score. If this procedure is employed, the mean differences between the two orders of response scales can be examined at both the item and test levels. The results are shown in Table 1. A systematic pattern of mean differences is evident. The positive response order tended to produce higher means on four out of five items compared to the reversed order. An overall multivariate test indicated that these mean differences between two forms of response order across five items were significant ($F=3.47$, $df=(5, 96)$, $p<.01$). Subsequent univariate t-tests indicated that, among the four mean

differences, two were significant ($p<.05$), one was marginally significant ($p<.053$), and the remaining one was not significant. When those effects on items were accumulated at the test level, the total score was significantly higher for the positive response order than for the reversed order. Those effects support the primacy effect hypothesis. No significant recency effect was found.

Insert Table 1 about here.

Factor Analysis

In performing factor analysis, researchers usually assign successive integers to response categories and obtain a Pearson correlation matrix. Results from this procedure will be compared later with those from the factor analysis using polychoric correlations. Pearson correlation matrices for each form of the PD scale is presented in the upper half of Table 2. Because the five items were designed to measure one common factor, i. e., personal distress, a one-factor model was fitted to the two Pearson correlation matrices separately. The parameters were estimated through the LISREL VI program (Joreskog & Sorbom, 1984). Path coefficients and residuals are presented in the left half of Table 3. The one-factor model fit the positive-form data very well (Chi-square = 2.82, df = 5, $p < .727$), while it was rejected as an inappropriate model for the reversed-form data (Chi-square = 24.01, df = 5, $p < .000$). These results were somewhat surprising because the internal consistency reliability

estimate within the positive-form data (alpha = .75) was lower than that within the reversed-form data (alpha = .82). After inspecting the discrepancies between observed and reproduced correlation matrices of form 2, it was found that the main source of poor fit was due to one data point, i.e., the correlation between item 4 and item 5. When the coefficients of the paths leading from the common factor to the five items are constrained to be equal between the two samples corresponding to the two forms of response scale, the whole model was rejected (Chi-square = 34.60, df = 20, p = .022). Consequently, the factor structure was not found to be equal between the two samples.

Insert Table 2 and Table 3 about here.

Based on the Pearson correlations, the above factor analysis assumes that the observed variables are measured by equal interval scales and are normally distributed. However, Likert-type scales are more reasonable treated as ordinal scales which measure an underlying continuous variable or trait. In this case, individuals' responses are jointly determined by the level of the latent trait and the threshold value of the response scales. The threshold values can be estimated from the inverse of the normal distribution function, and then polychoric correlations can be computed (Olsson, 1979). The inter-item polychoric correlations estimated by LISREL VI are presented in the lower half of Table 2. It was evident that the Pearson correlations were systematically lower than the corresponding polychoric

correlations. The threshold values, which are analogous to difficulty parameters in IRT, are presented in the left half of Table 5. It is interesting to note that the threshold values for the positive form were systematically lower than those for the reversed form, suggesting that the positive form was "easier" than the reversed form. As will be seen in the next paragraph, this result converges with the result from the IRT analysis. Finally, the factor structures based on polychoric correlations are shown in the right half of Table 3. It is not surprising to see that the factor loadings based on polychoric correlations were systematically higher than their corresponding factor loadings estimated from Pearson correlations. Because the chi-square value provided by LISREL VI is not correct with polychoric correlations, it was ignored. However, the other two goodness-of-fit measures, the Adjusted Goodness-of-fit Index and the Root Mean-square Residual, still suggested that the one-factor model fit the positive-form data fairly well while it did not fit the reversed-form data adequately.

IRT Analysis

Unidimensionality is a common assumption for most IRT models including the Graded Response model. This assumption might not be valid for the reversed-form data because it has been shown that the one-factor model did not fit this data set very well according to the Chi-square test. However, the Chi-square test is so sensitive to sample size that it is always significant with sufficiently large sample size. Therefore, traditional procedures for checking the unidimensionality assumption were also employed

here. Principal factor analysis was applied to the two data sets separately. Scree plots showed one "dominant" factor that influenced test performance within the two forms of scales respectively. For the positive-form data, the first eigenvalue (= 1.972) was 141 times the second eigenvalue (= 0.014). For the reversed-form data, the first eigenvalue (= 2.467) was 8 times the second eigenvalue (= .304). Consequently, the two sets of items were considered approximately unidimensional within each set.

Insert Table 4 and Table 5 about here.

Item and person parameters were estimated through the MULTILOG program (Thissen, 1986). The graded response model includes a discrimination parameter for each item and a difficulty parameter for each category boundary of each item. Since each of the five items employed in the present study had five response categories, there were four difficulty parameters for each item. Discrimination and difficulty parameters are presented in Table 4 and Table 5 respectively. The two sets of discrimination parameters have a Pearson r = .9794, while the two sets of difficulty parameters have a Pearson r = .8515. From these two tables, it can be seen that items with positive form of response scales are less discriminative and "easier" than those with reversed form of response scales. These results are consistent with the previous findings that positive-form items

had higher internal consistency and with the prediction of a primacy effect respectively. In responding to the positive-form items, subjects tended to choose the more positive response categories and obtained higher scores so that the estimated difficulty parameters of the positive-form items were systematically lower than those of the reversed-form items. This result was consistent with that from factor analysis. The estimated values of the two difficulty parameters of the fifth item for both forms can be ignored because their standard errors of estimate were extremely large due to the fact that few people chose the two most positive response categories of the last items in both forms, i.e., the two item were exetreme. Person parameters for the 102 people were also estimated. With the positive-form test, person parameters were estimated to have a mean of -0.024 and a standard deviation of 0.886, while with the reversed-form test, they were estimated to have a mean of -0.23 and a standard deviation of 0.950. The two means differed significantly from each other ($t = 2.81$, $p = .006$). This result implied that the two forms of tests produced different estimation of the latent trait of the same people. People tended to have a higher level of personal distress when their person parameters were estimated from the positive-form scale than when estimated from the reversed-form scale because the items in positive form were "easier" than their companions in reversed form as shown by the estimated difficulty parameters. The two sets of person parameters have a Pearson $r = .6835$.

Discussion

The present study found that the order of response scale labels significantly affected subjects' choices of the alternatives in Likert-type attitude scales. Subjects typically chose the first satisfactory alternative among the graded response categories so that a primacy effect was supported by present findings. This effect produced a higher mean for scales with positive order (from positive to negative end) than for those with reversed order at the item level and especially at the test level. In addition, the primacy effect altered the factor structures underlying both the Pearson and the polychoric correlation matrices of the two scale forms. A one-factor model fit the positive-form data very well while it was rejected for the reversed-form data, regardless of whether the Pearson or polychoric correlations were analyzed. The factor structures of the two dependent samples were evidently not equal. In addition, threshold values estimated by the LISREL VI program revealed that items in the positive form were "easier" than those in the reversed form, which was in agreement with the prediction of a primacy effect. Finally, analysis using an IRT model revealed that different forms of the same test changed the estimated values of item and person parameters. The response order effect disturbed the property of invariance of item and person parameters. Items with positive response order were estimated to be "easier" than those with reversed response order, while people responding to the items with positive response order were estimated to have "higher" levels of the latent trait than the

same people who responded to the items with reversed order. This phenomenon is consistent with the primacy effect.

Regarding the underlying structures of the correlation matrix, the present study only explored one model, i. e., a one-factor model assumed by the test constructor. However, there is always the possibility that more than one model may fit the same correlation matrix equally well. The invariance of the latent structure of another kind of model may be explored by future studies.

How is it possible that one-factor model fitted the positive-form data better than the reversed-form data while the positive-form items had lower internal consistency and discriminative ability than the reversed-form items? It is possible because the data points in a correlation matrix are not completely independent from one another. Given a 3-variable situation, for example, if r_{xy} and r_{yz} are known, then the range of r_{xz} can be determined as follows:

$$r_{xy}r_{yz} - \sqrt{(1-r_{xy}^2)(1-r_{yz}^2)} \leq r_{xz} \leq r_{xy}r_{yz} + \sqrt{(1-r_{xy}^2)(1-r_{yz}^2)}$$

Consequently, if one of the correlations in a matrix is peculiar (e.g., too high) then the one-factor model will fit the data poorly even though the estimated path coefficients, which reflect discriminative ability of each variable, are all high.

Item and test calibration using IRT models usually requires large samples. The sample size of the present study may be too small to produce an accurate estimation of the item parameters. In addition, five items were also insufficient to yield a stable estimation of person parameters. However, the patterns of the

results were revealing and consistent with the results from mean differences and factor analyses.

The present study avoids the confounding factor of sampling fluctuation by using repeated samples. In doing so, however, a different interfering factor, time, was introduced. Since no control group was employed, the attribution to primacy effect may be weakened by some rival hypotheses such as maturity. However, considering all available knowledge and theories, primacy effect is still the most reasonable factor to explain the systematic findings.

What is the practical implication of the present findings? Researchers using Likert-type scales routinely recode or flip the response scales of the negatively worded items. The present findings suggest that the recoded scores will differ from the raw scores obtained from the actual situation where the order of the response scales is reversed. In a longitudinal study, a group-comparison study, or a test-retest experiment of attitude change using Likert-type scales or other similar types of questionnaires as instruments, researchers may accidently shift the response scale order and confound its effects with interesting variables. Some researchers may even explicitly assume that the effect of reversing the response scale is negligible. The present study has suggested that the effect may affect the estimation of means, factor structures/ threshold values estimated by factor analysis, and item/person parameters estimated by an IRT model. Therefore, a primacy effect from this measurement factor needs to be

avoided, considered, held constant, or separated from the interested effects by cautious researchers.

References

Barsalou, L. W. (1982). Context-independent and context-dependent information in concepts. *Memory and Cognition, 10*, 82-93.

Bass, B. M., Cascio, W. F., and O'connor, E. J. (1974). Magnitude estimations of expressions of frequency and amount. *Journal of Applied Psychology, 59*, 313-320.

Becker, S. L. (1954). Why an order effect. *Public Opinion Quarterly, 18*, 271-278.

Belson, W. A. (1966). The effects of reversing the presentation order of verbal rating scales. *Journal of Advertising Research, 6*, 30-37.

Bishop, G. F., Oldendick, R. W., & Tuchfarber, A. J. (1985). The importance of replicating a failure to replicate: Order effects on abortion items. *Public Opinion Quarterly, 49*, 105-114.

Bradburn, N. M. & Miles, C. (1978). Vague quantifiers. *Public Opinion Quarterly, 43*, 92-101.

Brook, D. & Upton, G. J. G. (1974). Biases in local government elections due to position on the ballot paper. *Applied Statistics*, 23, 414-419.

Campbell, D. T. & Mohr, P. J. (1950). The effect of ordinal position upon responses to items in a check list. *Journal of Applied Psychology, 34*, 62-67.

Carp, F. M. (1974). Position effects on interview responses. *Journal of Gerontology, 29*, 581-587.

Chan, J. C. (1986). Sex, gender roles and empathy. Unpublished thesis. National Chengchi University, Taiwan, R. O. C.

Chase, C. I. (1969). Often is where you find it. *American Psychologist, 24,* 1043.

Churchill, G. A. and Peter, J. P. (1984). Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research, 21,* 360-375.

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *Catalog of Selected Documents in Psychology, 10,* 85.

Dodd, B. G. (1984). Attitude scaling: A comparison of the graded response and partial credit latent trait models. Doctoral dissertation, University of Texas at Austin.

Hakel, M. D. (1968). How often is often? *American Psychologist, 23,* 533-534.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,* 411-428.

Johnson, J. D. (1981). Effects of the order of presentation of evaluative dimensions for bipolar scales in four societies. *The Journal of Social Psychology, 113,* 21-27.

Joreskog, K. G. & Sorbom, D. (1984). LISREL VI: User's Guide (3rd ed.). Mooresville, IN: Scientific Software, Inc.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55,* 312-320.

Koch, W. R. (1983). Likert-type scaling using the graded response latent trait model. *Applied Psychological Measurement, 7*, 15-32.

Krosnick, J. A. & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51*, 201-219.

Mueller, J. E. (1970). Choosing among 133 candidates. *Public Opinion Quarterly, 34*, 395-402.

Mathews, C. O. (1929). The effect of the printed response words on an interest questionnaire. *Journal of Educational Psychology, 30*, 128-134.

McClendon, M. J. (1986). Response-order effects for dichotomous questions. *Social Science Quarterly, 67*, 205-211.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44*, 443-460.

Ortony, A., Schallert, D. L., Reynolds, R. E., and Antos, S. J. (1978). Interpreting metaphors and idioms: Some effects of context on comprehension. *Journal of Verbal Learning and Verbal Behavior, 17*, 465-477.

Parducci, A. (1968). Often is often. *American Psychologist, 23*, 828.

Payne, S. L. (1951). The Art of Asking Questions. Princeton: Princeton University Press.

Payne, J. D. (1972). The effects of reversing the order of verbal rating scales in a postal survey. *Journal of Market Research Society, 14*, 30-44.

Pepper, S. & Prytulak, L. S. (1974). Sometimes frequently means seldom: Context effect in the interpretation of quantitative expressions. *Journal of Research in Personality, 8*, 95-101.

Powers, E. A., Morrow, P., Goudy, W. J., & Keith, P. M. (1977). Serial order preference in survey research. *Public Opinion Quarterly, 41*, 80-85.

Roth, E. M. & Shoben, E. J. (1983). The effect of context on the structure of categories. *Cognitive Psychology, 15*, 346-378.

Rugg, D. and Cantril, H. (1944). The wording of questions. In *Gauging Public Opinion* (H. Cantril, ed.), Chapter 2. Princeton: Princeton University Press.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, No.17.

Sauber, M. H. (1983). Impact of non-stimulus object factors on rating scale responses. Unpublished dissertation, University of Texas at Austin.

Schertzer, C. B. & Kerman, J. B. (1985). More on the robustness of response scales. *Journal of the Market Research Society, 27*, 261-282.

Schuman, H. & Presser, S. (1981). Questions and Answers in Attitude Surveys. New York: Academic Press.

Simon, H. A. (1957). Models of Man. New York: Wiley.

Stouffer, S. A. & DeVinney, L. C. (1949). how personal adjustment varied in the army--By background characteristics of the soldiers. In S. A. Stouffer, E. A. Schuman, L. C. Devinney, S. A. Star, and R. M. Williams (eds.), *The American Soldiers:*

Adjustment During Army Life. Princeton: Princeton University Press.

Thissen, D. & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin, 104*, 385-395.

Tourangeau, R. & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103*, 299-314.

Wildt, A. R. & Mazis, M. B. (1978). Determinants of scale response: label versus position. *Journal of Marketing Research, 15*, 261-267.

Worcester, R. M. & Burns, T. R. (1975). A statistical examination of the relative precision of verbal scales. *Journal of the Market Research Society, 17*, 181-196.

Wright, B. D. (1967). Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational conference on Testing Problems. Princeton, NJ: Educational Testing Service, 1986.

Table 1. Mean differences of item and total scores between two orders of response scales.

Item	Mean		SD		t-value	2-tail p
	Pos.*	Rev.*	Pos.	Rev.		
1.	2.088	1.873	1.321	1.287	1.96	.053
2.	2.284	1.667	1.367	1.269	4.13	.000
3.	1.873	1.608	1.240	1.228	2.09	.039
4.	1.128	1.029	1.096	1.121	.90	.368
5.	.584	.614	.840	.927	-.34	.734
Total	8.000	6.832	4.198	4.470	3.31	.001

* Pos. = positive form; Rev. = Reversed form.

Table 2. Pearson and polychoric correlations among five items for two orders of response scales.

Item	1.	2.	3.	4.	5.
Pearson Correlations					
1.		.277	.629	.491	.392
2.	.450		.349	.220	.227
3.	.765	.530		.515	.377
4.	.585	.439	.567		.404
5.	.343	.138	.348	.547	
Polychoric Correlations					
1.		.323	.674	.546	.455
2.	.455		.368	.238	.281
3.	.814	.566		.553	.418
4.	.675	.487	.602		.493
5.	.444	.141	.391	.644	

Note: The inter-item correlations are shown above the diagonal for the positive form and below the diagonal for the reversed form.

Table 3. Path coefficients and residuals estimated by the one-factor model for different scale forms and different correlations from Table 2.

	For Pearson Corr.				For Polychoric Corr.			
	Pos. Form		Rev. Form		Pos. Form		Rev. Form	
Item	LV	Res.	LV	Res.	LV	Res.	LV	Res.
1.	.769	.408	.856	.268	.811	.343	.906	.179
2.	.394	.845	.569	.676	.418	.825	.564	.682
3.	.805	.352	.878	.230	.811	.343	.881	.223
4.	.649	.579	.691	.523	.692	.521	.747	.442
5.	.514	.736	.436	.810	.576	.668	.510	.740
AGFI	.984		.883		.968		.799	
RMR	.026		.074		.034		.086	

Note: Pos.= Positive; Rev.= Reversed; LV = Latent Variable. Res.= Residuals; AGFI = Adjusted Goodness of Fit Index; RMR = Root Mean-square Residual.

Table 4. Estimated item discrimination parameters and standard errors from the graded response model.

	Positive Form		Reversed Form	
Item	a	SE	a	SE
1	2.550	.411	3.157	.545
2	.829	.247	1.183	.247
3	2.504	.399	3.011	.602
4	1.639	.330	1.916	.393
5	1.170	.308	1.059	.300

Note: a = Discrimination parameters; SE = Standard error of estimates.

Table 5. Estimated threshold values from factor analysis and item difficulty parameters from the graded response model.

	Factor Analysis			The Graded Response Model				
Item	T1	T2	T2-T1	b1	(SE)	b2	(SE)	b2-b1
1.	-1.19	-1.13	.06	-1.427	(.230)	-1.289	(.184)	.138
	-.27	-.06	.21	-.411	(.141)	-.133	(.137)	.278
	.25	.37	.12	.223	(.151)	.403	(.138)	.207
	.86	1.09	.23	1.024	(.190)	1.291	(.194)	.087
2.	-1.05	-.88	.17	-2.401	(.806)	-1.589	(.425)	.812
	-.51	.06	.57	-1.185	(.463)	.050	(.255)	1.235
	-.02	.62	.64	-.106	(.348)	1.041	(.362)	1.147
	.75	1.18	.43	1.665	(.609)	2.141	(.556)	.476
3.	-1.24	-1.00	.24	-1.503	(.228)	-1.160	(.170)	.343
	-.05	.24	.29	-.149	(.148)	.218	(.144)	.367
	.46	.56	.10	.502	(.160)	.606	(.155)	.104
	1.09	1.35	.26	1.340	(.233)	1.671	(.264)	.331
4.	-.38	-.26	.12	-.622	(.214)	-.424	(.184)	.198
	.49	.62	.13	.600	(.226)	.753	(.201)	.153
	1.05	1.29	.24	1.482	(.313)	1.742	(.362)	.260
	2.06	1.56	-.50	3.108	(.720)	2.242	(.493)	-.866
5.	.27	.21	-.06	.417	(.269)	.393	(.309)	-.024
	1.05	1.18	.13	1.836	(.471)	2.260	(.634)	.424
	1.76	1.65	-.11	3.244	(1.109)	3.276	(.923)	.032
	*	1.89	*	11.031	(480.79)	3.861	(1.17)	-7.17

Note: T1 (or T2) = Threshold values for the positive (or reversed) form; b1 (or b2) = Difficulty parameters for the positive (or reversed) form; SE = Standard error of estimates. * = No people responded to the highest category, so that number of threshold is reduced for this item.